# Accurate Human Pose Estimation using RF Signals

Chunyang Xie†, Dongheng Zhang¶, Zhi Wu¶, Cong Yu†, Yang Hu‡, Qibin Sun¶, and Yan Chen¶

†School of Information and Communication Engineering, University of Electronic Science and Technology of China
Emails: {chunyangxie, congyu}@std.uestc.edu.cn
¶School of Cyber Science and Technology, University of Science and Technology of China
Emails: dongheng@ustc.edu.cn, wzwyyx@mail.ustc.edu.cn, {qibinsun, eecyan}@ustc.edu.cn
‡School of Information Science and Technology, University of Science and Technology of China
Emails: eeyhu@ustc.edu.cn

***Abstract*—Radio-frequency (RF) based human sensing technologies, due to their great practical value in various applications and privacy-preserving nature, have gained tremendous attention in recent years. However, without fully exploiting the characteristics of radio signals, the performance of existing methods are still limited. First, RF features of the moving human body have different representations in dimensions such as channel and scale, which is challenging when performing feature fusion. Besides, the human body is specularly reflective with respect to the radar, which means the human body cannot be fully captured by a single RF snapshot. Therefore, the radar signal reflected by the human body is sparse and incomplete, which is difficult to extract high-quality features for 3D human pose estimation. In this paper, we present the RF-based Pose Machines (RPM), a novel framework which can generate 3D skeletons from RF signals. Considering the characteristics of RF signals, RPM includes several modules to overcome the challenges. Firstly, a Multidimensional Feature Fusion (MFF) backbone is designed to effectively fuse radio signals based on the channels' correlation and maintain high-quality feature via a multi-scale fusion block. A Spatio-Temporal Attention network is then designed to reconstruct 3D skeletons by modeling the non-local spatio-temporal relationships. To evaluate the performance of our RPM framework, we construct a large-scale dataset of synchronized 3d skeletons and RF signals, RFSkeleton3D. Our experimental results show that RPM locates 3D key points of the human body with an average error of $5.71cm$ and maintains its performance in new environments with occlusion or bad illumination. The dataset and codes will be made in public.**

***Index Terms*—RF sensing, 3D Human Pose Estimation, Deep Learning, Smart Homes.**

## I. Introduction

HUMAN pose estimation (HPE) has drawn increasing attention due to its various application scenarios such as smart homes, enterprise security, virtual reality, and medical care. The 3D HPE based on visual images or videos has made tremendous progress with the recent advances of deep learning. However, capturing 3D human skeletons via cameras still has some drawbacks. Firstly, camera-based 3D HPE methods can not be well adapted to complex scenarios in practice, such as occlusion, bad illumination, blurry, etc. In addition, privacy issues cannot be ignored, especially in smart home and health care application scenarios, where the extensive use of cameras can bring privacy concerns.

Considering the aforementioned drawbacks, significant research efforts have been spent on advanced sensing technologies, which aim to perceive and understand human activities through radio signals. Recent pioneer studies [1], [2] demonstrate that RF signals carry an impressive amount of information about people which can be used to generate 2D and 3D skeletons of human bodies. However, the representations of the reflected RF signals have different distributions in terms of channel and scale, leading to the difficulty of feature fusion. Furthermore, a single RF snapshot cannot capture the whole human body due to the specularity of the human body with respect to RF, which leads to the sparsity and incompleteness. The above limitations make RF-based human pose estimation a challenging task.

**Our contributions:** In this paper, we propose RF-based Pose Machines (RPM), to locate and reconstruct accurate 3D human skeletons from RF signals. Considering the characteristics of RF signals, RPM includes several modules to overcome a series of challenges. Firstly, an effective Multi-dimensional Feature Fusion (MFF) backbone network is designed to combine horizontal and vertical RF features with channel-wise attention and maintain scale-insensitive feature representations via a deformable multi-stage convolution module. Then, a Spatio-Temporal Attention (STA) network is proposed to construct 3D skeletons with multi-head attention from the sparse and incomplete RF features, in which the spatial attention module is designed to model non-local joint relationships and the temporal attention module is designed to refine 3D skeleton sequences with temporal coherency.

To evaluate the proposed RPM framework, we collect a large-scale dataset of synchronized 3D skeletons and RF signals, named RFSkeleton3D. The evaluation results show that RPM locates key points of the human body with an average error of $5.71cm$, achieving state-of-the-art accuracy, and maintains its performance in new environments with occlusion or bad illumination.

## II. Related Work

### A. Camera-based 3D Human Pose Estimation

Camera-based 3D HPE methods aim to estimate 3D human joint locations from images or videos. Recently, with the devel-

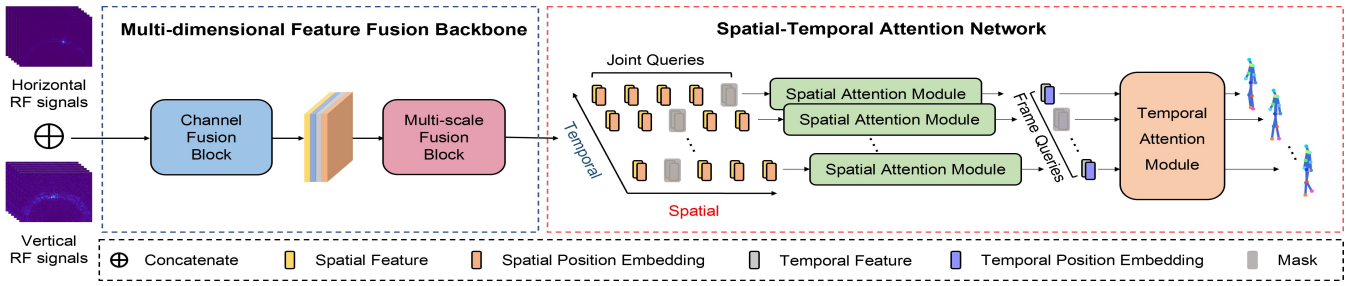


Fig. 1. The proposed RPM framework for RF-based 3D human pose estimation. Horizontal and vertical RF siganls are concatenated as input to the MFF backbone, in which features are fused cross different channels and feature scales. Then Spatial Attention Module models non-local skeleton relationships from input joint queries with masks to construct body parts that cannot be captured by RF snapshots. Temporal Attention Module refines 3D poses based on temporal coherency learned from partially masksed frame queries.

opment of deep learning, CNN-based solutions have improved the pose estimation performance significantly. Specifically, Pavlakos et al. [3] proposed a volumetric representation for 3D human pose and predicted 3D pose from coarse to fine. Zhou et al. [4] predicted 3D poses by maintaining 2D heatmaps and depth regression. Some methods attempt to obtain a 3D skeleton from 2D poses produced by 2D HPE methods. Choi et al. [5] proposed a graph convolutional neural network-based system to estimate 3D pose from the 2D pose. Hybrik [6] transformed 3D pose to mesh and refined the mesh estimation and pose estimation at the same time.

Despite impressive performance achieved by camera-based approaches, the accuracy and robustness of these methods can be impaired by occlusion, bad illumination, and blurry. Furthermore, privacy issues cannot be ignored when cameras are deployed to monitor human subjects. The RPM framework, on the other hand, achieves the HPE through RF signals, which breaks the limitation of existing camera-based HPE methods.

### B. Wireless Sensing

Considering the aforementioned drawbacks of camera-based approaches, significant research efforts [7]–[9] have been spent on advanced sensing technologies via radio signals. WiPose [10] encoded skeleton prior knowledge and took 3D velocity profile as input to generate 3D skeletons with an RNN network. Wang et al. [11] constructed CSI images that contain both pose and position information and designed a neural network to predict 3D pose from CSI features.

Compared with WiFi, RF signals can achieve better resolution due to the larger bandwidth and more antennas. Zhao et al. [1], [2] proposed a neural network to predict 2D/3D human pose via RF signals. Sengupta et al. [12] proposed a network, mm-Pose, to estimate 3D human pose via point cloud data from mmWave radars.

In this paper, we propose RPM, a novel framework to estimate the 3D human skeleton with mmWave radars. RPM consists of a powerful Multi-dimensional Feature Fusion backbone network to extract scale-insensitive high-resolution feature representations, and a Spatio-Temporal Attention model to model non-local spatio-temporal relationships. Without bells and whistles, our RPM achieves state-of-the-art performance in the 3D HPE task.

## III. The Proposed RPM

### A. Overview

The overview of the proposed RPM framework is shown in Fig. 1. Horizontal and vertical RF signals are concatenated and fed into the proposed MFF backbone to maintain high-quality feature vectors. Next, a Spatial Attention Module (SAM) models non-local skeleton relationships to predict "hard" body parts using masked feature vectors. Finally, a Temporal Attention Module (TAM) outputs refined 3D skeletons based on feature sequences with frame-wise masks. We describe each component in detail as below.

### B. Multidimensional Feature Fusion backbone

To achieve RF-based human pose estimation, we need to extract fine-grained features from RF signals. Horizontal RF heatmaps are projections of RF signals on a plane parallel to the ground, which can provide the human body position, whereas vertical RF heatmaps are projections of RF signals on a plane perpendicular to the ground, which contains information about body parts. Considering the interdomain differences of RF signals, we propose a Multidimensional Feature Fusion backbone network to effectively fuse features at different scales and channels in the horizontal and vertical directions. First, a channel fusion block, extracts channel-wise fused features from concatenated RF signals. Next, considering the scale difference of horiziontal and vertical RF signals, we design a multi-scale fusion block to adaptively adjust the size of the receptive field, in which deformable convolution layers are applied to a multi-resolution feature extraction network motivated by [13]. Finally, a light weight multilayer perceptron (MLP) network is adopted as a classification head to transform the extracted feature into vector which will be fed to the Spatio-Temporal Attention network for the regression task.

### C. Spatial-Temporal Attention network

Unlike a camera that captures the projection of all unoccluded body parts, mmWave radar can only receive the reflected signal of a subset of limbs at each RF snapshot. We propose a Spatio-Temporal Attention network to reconstruct 3D skeletons from RF signals with information about an unknown subset of the body parts. As shown in Fig. 2, the STA

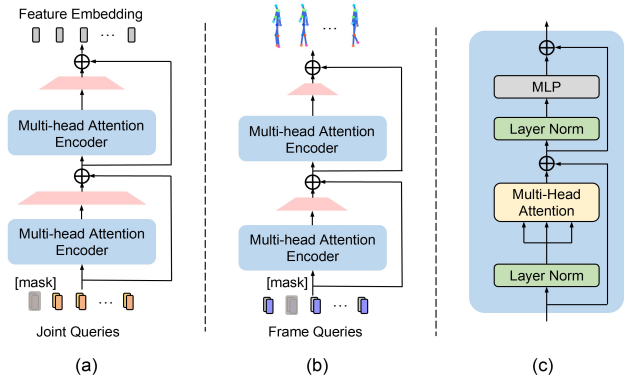


Fig. 2. Architecture of Spatial-Temporal Attention Network: (a) Spatial Attention Module architecture, (b) Temporal Attention Module architecture, (c) Multi-head Attention Encoder architecture. Linear layers are inserted between adjacent Multi-head Attention Encoders to reduce feature dimension progressively.

network consists of two modules: Spatial Attention Module (SAM) and Temporal Attention Module (TAM).

*1) Spatial Attention Module:* Since signal from a single RF snapshot only has limited information, the goal of the spatial attention module is to recover the remaining body parts by modeling non-local skeleton relationships via the multi-head self-attention mechanism, which includes the following three steps.

**Patch Embedding.** Assume the feature from the MFF backbone is $\boldsymbol{X}_{MFF} \in \mathbb{R}^{F\times K\times C}$, where $F$ is the length of the video sequence, $K$ is the length of the output feature vector, and $C$ is the dimension. We use a trainable linear projection layer to embed each vector to a feature $\boldsymbol{X}_{embed} \in \mathbb{R}^{F\times J\times C}$, where $J$ is the number of 3D joints.

To simulate the specularity of human body with respect to RF signals, we design a Masked Joint Modeling (MJM) to fully activate the bi-directional attention in the multi-head attention encoder, which masks some input queries at random. Given the embedded feature of each frame $\boldsymbol{x} \in \mathbb{R}^{J\times C}$, the output feature $\boldsymbol{H} \in \mathbb{R}^{J\times C}$ can be written as:

$$\boldsymbol{H} = [\boldsymbol{x}^1; \cdots; \mathrm{MJM}(\boldsymbol{x}^i); \cdots; \boldsymbol{x}^J] + \boldsymbol{E}_{SPE} \tag{1}$$

where $\boldsymbol{E}_{SPE} \in \mathbb{R}^{J\times C}$ is the learnable position embedding for each joint.

We enforce the encoder to regress spatial feature embedding of all joints from the masked queries, which is in a spirit similar to simulating scenarios where only a subset of body parts can be captured by RF signals. As a result, the multi-head attention encoder will consider spatial features of other joints for modeling better human skeleton, which encodes non-local information.

**Multi-head Self-attention.** Given the embedded feature of a single frame $\boldsymbol{H}$, we first compute a query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$ and value matrix $\boldsymbol{V}$ by linear transformations $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$:

$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{K} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{V} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{V}}. \tag{2}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{J\times C}$, $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{C\times C}$.

The scaled dot-producted attention can be written as:

$$\mathrm{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{C}}\right)\boldsymbol{V} \tag{3}$$

Furthermore, the multi-head attention is the concatenation of the attention in Eqn. (3) computed by $h$ attention heads.

$$\mathrm{MHAttn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}\left(\boldsymbol{O}_1, \cdots, \boldsymbol{O}_h\right)\boldsymbol{W}_{out} \tag{4a}$$

$$\boldsymbol{O}_i = \mathrm{Attn}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right), \quad i \in [1, 2, \ldots h] \tag{4b}$$

where $\boldsymbol{W}_{out} \in \mathbb{R}^{C\times C}$ is the projection matrix.

We can compute the self-attentive feature to model non-local dependencies of different joints with L-layer multi-head attention encoder, which can be presented as follows:

$$\boldsymbol{H}'_l = \mathrm{MHAttn}(\mathrm{LayerNorm}(\boldsymbol{H}_{l-1}) + \boldsymbol{H}_{l-1} \tag{5a}$$

$$\boldsymbol{H}_l = \mathrm{MLP}(\mathrm{LayerNorm}(\boldsymbol{H}'_l)) + \boldsymbol{H}'_l, \quad l = 1, 2, \ldots L \tag{5b}$$

$$\boldsymbol{Y} = \mathrm{LayerNorm}(\boldsymbol{H}_L) \tag{5c}$$

where $\boldsymbol{H}_0 = \boldsymbol{H}$ is the initial embedded feature, the $\mathrm{LayerNorm}(\cdot)$ is the layer normalization [14].

**Dimensionality Reduction.** Our task is to regress the 3D coordinates of the human skeleton from the RF features. It is not appropriate to apply a multi-head attention encoder directly, since the dimensionality of the hidden embedding remains constant. Therefore, our Spatial Attention Module (SAM) applies a progressive dimensionality reduction scheme, i.e., adding a multi-head attention encoder for self-attention and a learnable linear layer for dimensionality reduction in an alternating manner. For a module with a depth of $N$, the final output is

$$\boldsymbol{Z}_n = \mathrm{Linear}(\boldsymbol{Y}_n) + \mathrm{Linear}(\boldsymbol{Z}_{n-1}), \quad n = 1, 2, \ldots N \tag{6}$$

where $\boldsymbol{Y}_n$ is the output of the multi-head attention encoder.

*2) Temporal Attention Module:* It is difficult to generate accurate 3D skeleton based on information from a single frame. Therefore we propose a Temporal Attention Module (TAM) to model temporal dependencies across the 3D skeleton sequences. The TAM shares the same design philosophy as the SAM, which also includes several multi-head attention encoders followed by a linear layer for dimension reduction. One of the major differences is the input. For the output feature embedding of SAM $\boldsymbol{X}_{spatial} \in \mathbb{R}^{F\times J\times C'}$, the input feature embedding of TAM is $\boldsymbol{X}_{temporal} \in \mathbb{R}^{F\times(J\cdot C')}$, in which all spatial features of each frame are concatenated to a vector $\boldsymbol{x}_t \in \mathbb{R}^{1\times(J\cdot C')}$. The other change is the Masked Frame Modeling (MFM). Unlike MJM, we masked the input queries of random frames to make the multi-head encoder to predict all 3D skeleton sequences, which enforces the module to facilitate non-local interactions in the time dimension. The procedure can be formulated as:

$$\boldsymbol{X}_{temporal} = [\boldsymbol{x}_t^1; \cdots; \mathrm{MFM}(\boldsymbol{x}_t^i); \cdots; \boldsymbol{x}_t^F] + \boldsymbol{E}_{TPE} \tag{7}$$

where $\boldsymbol{E}_{TPE} \in \mathbb{R}^{F\times(J\cdot C')}$ is the learnable position embedding for each frame.

TABLE I

THE MPJPE METRIC (UNIT: $mm$) COMPARISONS ON RFSKELETON3D DATASET AMONG STATE-OF-THE-ART METHOD AND OUR FRAMEWORK. BOLD FONTS REPRESENT THE BEST RESULTS.

| Method | Nose | Neck | Sho | Elb | Wri | Hip | Knee | Ank | Mean (↓) |
|---|---|---|---|---|---|---|---|---|---|
| RFPose3D [2] | 81.4 | 52.0 | 90.1 | 120.4 | 135.1 | 89.9 | 144.9 | 167.4 | 116.3 |
| mm-Pose [12] | 165.8 | 67.3 | 203.3 | 233.8 | 261.6 | 138.6 | 162.3 | 169.9 | 183.7 |
| RPM (ours) | **57.5** | **37.2** | **49.1** | **64.9** | **68.2** | **46.5** | **58.1** | **65.1** | **57.1** |

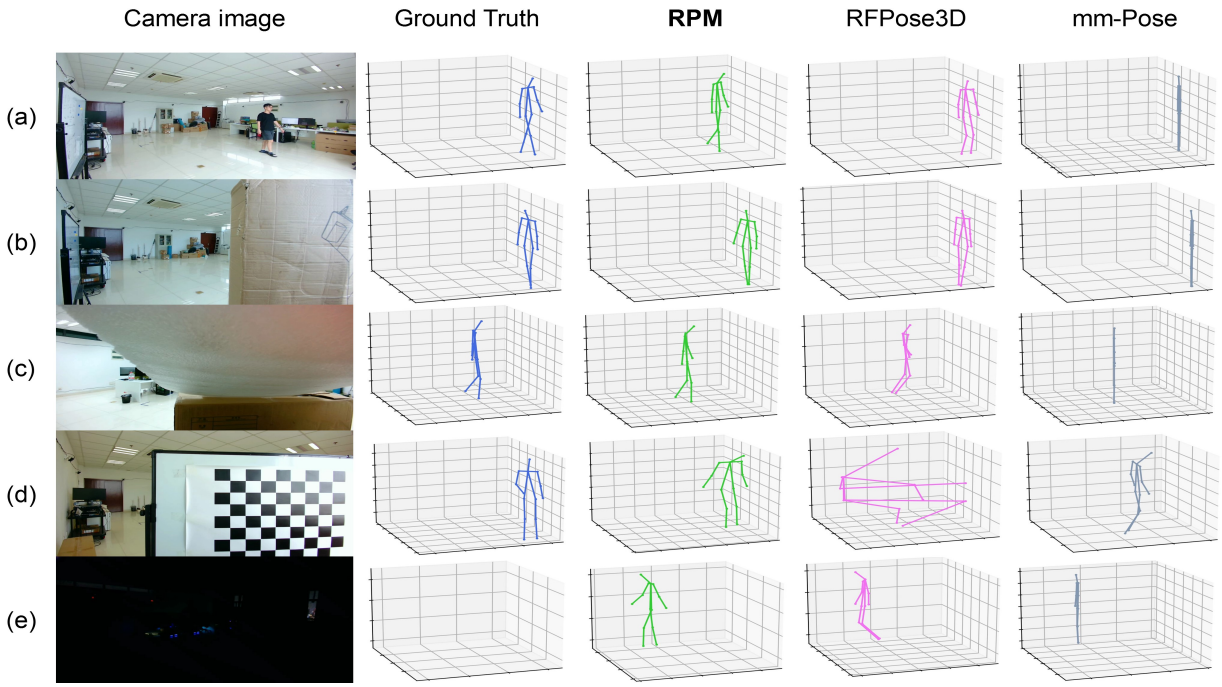


Fig. 3. Qualitative results under various scenarios: (a) basic environment, (b) occluded by a carton (c) occluded by foam, (d) occluded by a whiteboard, (e) low illumination. Note that the ground truth under occlusion is only of reference because some cameras are also occluded. Besides, the multi-camera system cannot generate ground truth in dark scenes.

### D. Loss Function

Let $\hat{\boldsymbol{P}}$ denote the ground truth 3D joint locations, $\boldsymbol{P}$ denote the 3D pose predictions, $\boldsymbol{P}^{root}$ and $\hat{\boldsymbol{P}}^{\boldsymbol{root}}$ are the coordinates of the center point of the human torso calculated from predicted and ground truth 3D poses, respectively. The training procedure aims to minimize the Euclidean or $L2$ distance between predictions and labels for all skeletons of each sequence. The location loss function is defined as:

$$\mathcal{L}_{loc} = \frac{1}{F}\sum_{i=1}^{F}\left\|\boldsymbol{P}_i^{root} - \hat{\boldsymbol{P}}_{\boldsymbol{i}}^{\boldsymbol{root}}\right\|_2 \tag{8}$$

where $F$ is the sequence length. And the pose estimation loss function is the average $L2$ distance between normalized predicted 3D skeletons and ground truth skeletons, which is defined as:

$$\mathcal{L}_{pose} = \frac{1}{FJ}\sum_{i=1}^{F}\sum_{k=1}^{J}\left\|(\boldsymbol{P}_i^k - \boldsymbol{P}_i^{root}) - (\hat{\boldsymbol{P}}_{\boldsymbol{i}}^{\boldsymbol{k}} - \hat{\boldsymbol{P}}_{\boldsymbol{i}}^{\boldsymbol{root}})\right\|_2 \tag{9}$$

where $J$ is the number of joints.

Our overall objective function is written as:

$$\mathcal{L} = \mathcal{L}_{loc} + \mathcal{L}_{pose} \tag{10}$$

## IV. EVALUATION

We first evaluate the performance of RPM and existing state-of-the-art methods [2], [12] on the RFSkeleton3D dataset. Then we demonstrate the effectiveness of each module through ablation experiments and provide insights for the non-local interactions and temporal attention.

### A. Experimental Setup

**Baseline:** RFPose3D [2] is the state-of-the-art RF-based 3D HPE method. We implement the RFPose3D model as our baseline, which regards the 3D HPE task as a joint classification problem. Besides, we also implement mm-Pose [12] as the other baseline method.

**Data:** Inspired by prior work [2], we implement a portable and robust data collection testbed, in which two Frequency Modulated Continuous Wave (FMCW) radars are adopted to capture horizontal and vertical RF signals and a multi-camera system is designed to obtain ground truth 3D skeletons by

TABLE II
ANALYSIS ON DIFFERENT BACKBONES.

| Backbone | MPJPE (↓) |
|---|---|
| ResNet18 | 65.4 |
| ResNet50 | 59.7 |
| MFF backbone (ours) | **57.1** |

TABLE III
ANALYSIS ON SPATIAL-TEMPORAL ATTENTION NETWORK.

| Method | SAM | TAM | MPJPE (↓) |
|---|---|---|---|
| RPM | × | × | 99.2 |
| | ✓ | × | 75.5 |
| | ✓ | ✓ | **57.1** |

performing triangulation on 2D poses. We randomly ask 9 volunteers to collect data in 10 scenarios such as rooms with different shading or lighting conditions so that the network cannot learn the correlation between the subject and the scenario. The dataset, which is named RFSkeleton3D, contains $174050$ 3D skeleton frames and $348100$ RF frames. We split the RFSkeleton3D dataset into two parts: the training dataset and the validation dataset. The training dataset includes $80\%$ of RF data. The validation dataset includes the remaining RF frames. All models are evaluated on the same training and validation sets. The length of the input RF sequences is 20 frames. The sliding window method with a window length of 10 is applied for data augmentation during the training procedure.

**Evaluation metric:** We use the most common evaluation metric Mean Per Joint Position Error (MPJPE) to evaluate the performance of RPM and baseline methods. MPJPE is a metric for evaluating 3D human pose, which measures the Euclidean distances between the ground truth joints and the predicted joints.

**Implement Details:** We train the baseline methods and our RPM on 2 NVIDIA V100 GPUs. For RPM, the batchsize is 16. We adopt cosine annealing with warmup 40 epochs as the learning rate scheduler. The maximum learning rate is $3 \times 10^{-4}$. The total number of training epochs is 200.

### *B. Results*

**General Performance:** Table I shows the performance of two baseline methods and our RPM framework. Several important observations can be obtained from Table I. First of all, our RPM performs significantly better than the baseline methods, achieving $57.1mm$ while RFPose3D and mm-Pose achieve $116.3mm$ and $183.7mm$, respectively. Secondly, all of the methods show some degradation in performance when estimating joint points with large motion amplitudes such as elbow, wrist, and ankle. This may be due to the small size and rapid movement of these body parts, making it difficult for radar to accurately locate them.

**Performance Under Occlusion or Low Illumination:** Occlusion and low illumination are very challenging for traditional camera-based 3D HPE methods. Nevertheless, the characteristics of RF signals make it possible to estimate 3D human poses accurately. In these challenging cases, quantitative evaluation cannot be conducted since multi-camera systems cannot produce accurate labels. Therefore, we only show the qualitative results. As shown in Fig.3, RPM works well in such scenarios, which demonstrates the advantage of our framework over the camera-based solutions. It is worth noting that the performance of all methods are severely affected when a large metal whiteboard is used to obscure the mmWave radar. In the fourth row, we observe that RFPose3D does not produce a reasonable human skeleton, and mm-Pose also produces substantial deviations. However, our RPM, thanks to powerful network design and the novel spatial-temporal attention mechanism, can still produce a reasonable human skeleton within the error tolerance.

TABLE IV
ANALYSIS ON HYPERPARAMETERS.

| Input Length | 10 | 14 | 20 | 30 | 40 |
|---|---|---|---|---|---|
| MPJPE (↓) | 61.5 | 60.3 | **57.1** | 60.1 | 61.1 |
| MJM Percentage | 10% | 20% | 30% | 40% | 50% |
| MPJPE (↓) | 58.2 | 61.9 | 57.9 | **57.1** | 59.1 |
| MFM Percentage | 0% | 10% | 20% | 30% | 40% |
| MPJPE (↓) | 57.3 | **57.1** | 57.9 | 59.1 | 59.7 |

### *C. Ablation Study*

To evaluate the contribution of the individual components of our RPM and the impact of hyperparameters, we conduct extensive ablation experiments on the RFSkeleton3D dataset.

**Analysis on Model Design.** We first study the behavior of different CNN backbone architectures. ResNet [15] is picked for this experiment. In Table II, we observe that our framework achieves a competitive performance of $59.7mm$ even when using a pre-trained ResNet50 backbone. When changing the backbone to our proposed MFF network, we observe further improvement, achieving $57.1mm$, which demonstrates the effectiveness of the MFF backbone.

We also conduct an ablation study on our Spatial-Temporal Attention network. We implement a modified network that includes only the MFF backbone and several linear layers to demonstrate the effect of the STA network. Furthermore, we also implement a network in which only the Spatial Attention Module (SAM) is followed by the MFF backbone to evaluate the effect of the Spatial Attention Module (SAM).

The output results are regressed into 3D poses via a small regression head which includes 2 linear layers. As shown in Table III, we observe degradation in the performance of the network after removing the Spatial-Temporal Attention network, but it still yields competitive results ($99.2mm$).

When the SAM is combined, the overall performance of RPM is improved, achieving $75.5mm$, which demonstrates that the non-local joint relationships are beneficial to 3D HPE task. Since TAM refines 3D skeletons based on temporal attention, RPM's performance is further improved after adding TAM, achieving $57.1mm$.

**Analysis on Architecture Parameters.** We explore the various hyperparameter settings to find the optimal architecture

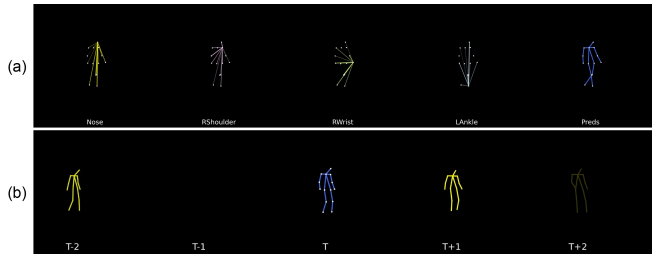


Fig. 4. Visualization of spatio-temporal attention in RPM: (a) spatial attention, (b) temporal attention. We visualize the spatial attention between specific joints and all other joints. Note that the temporal attention shown is relative to the human body at the moment of T (the blue skeleton in the middle). Brighter color indicates stronger attention.

parameters. First of all, we report the ablation experiment results regarding the RF sequence length. From the first sub-table in Table IV we can observe that when the sequence length is 20 frames, which is actually the data collected during 1 second by mmWave radar, the performance is the best. Too long or too short RF sequences will lead to performance degradation.

To better model non-local relationships in spatial and temporal domains, we propose two technologies, which are named Masked Joint Modeling (MJM) and Masked Frame Modeling (MFM). We also report the ablation experiment results of these two technologies. As shown in the second sub-table in Table IV, when we gradually increase the MJM ratio from $0\%$ to $40\%$, the performance of RPM improves. However, when the ratio exceeds $50\%$, the performance of the model starts to degrade. This is because too many missing spatial queries caused by the specularity of the human body with respect to RF and the artificial masking would increase the learning difficulty of our model. Similarly, from the third sub-table in Table IV, we can observe that RPM achieves the best performance when $10\%$ temporal queries are randomly masked.

### *D. Attention Visualization*

To further understand the effect of RPM in learning spatial-temporal interactions among joints and RF sequences, we visualize spatial and temporal attention in our model.

Fig. 4 shows the visualization of the spatio-temporal attention. The brighter color indicates stronger interaction. As shown in the first row in Fig. 4, RPM tends to predict specific joints based on relevant non-local joints. For example, RPM attends to the upper parts of the human body when predicting joints like the nose and shoulder. For the wrist position prediction, however, RPM attends to both upper and lower parts like the shoulder, elbow, and knee. It is reasonable that those joints provide strong cues to the 3D pose and subsequently the wrist position.

Furthermore, we visualize the temporal attention the temporal attention relative to a specific moment of $T$ (the middle blue skeletons). As indicated in the second row in Fig. 4, RPM attends more to the feature at $T-2$ and $T+1$ moments. Note that RPM focuses on features at different moments in different sequences. This is because the movement of the human body changes at different moments. RPM adaptively models temporal dependencies based on specific sequences.

## V. Conclusion

In this paper, we proposed RPM, a novel framework for accurately reconstructing the 3D human skeleton using RF signals. RPM consists of a Multi-dimensional Feature Fusion backbone network to extract scale-insensitive high-resolution feature representations from horizontal and vertical RF signals, a Spatio-Temporal Attention network to model non-local spatio-temporal 3D skeleton relationships. Furthermore, we constructed a large-scale dataset of synchronized 3d skeletons and RF signals, named RFSkeleton3D, to facilitate further research in RF sensing. The experimental results demonstrated that RPM could locate 3D key points of the human body with an average error of $5.71cm$ and maintain its performance in new environments with occlusion or bad illumination.

## References

[1] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[2] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 267–281.

[3] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7025–7034.

[4] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 398–407.

[5] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *European Conference on Computer Vision (ECCV)*, 2020.

[6] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383–3393.

[7] D. Zhang, Y. Hu, and Y. Chen, "Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3904–3914, 2020.

[8] C. Qiu, D. Zhang, Y. Hu, H. Li, Q. Sun, and Y. Chen, "Radio-assisted human detection," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[9] G. Zhang, D. Zhang, Y. He, J. Chen, F. Zhou, and Y. Chen, "Multi-person passive wifi indoor localization with intelligent reflecting surface," *arXiv preprint arXiv:2201.01463*, 2022.

[10] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3d human pose construction using wifi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[11] Y. Wang, L. Guo, Z. Lu, X. Wen, S. Zhou, and W. Meng, "From point to space: 3d moving human pose estimation using commodity wifi," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2235–2239, 2021.

[12] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.